ROLE OF MULTIVARIATE DATA ANALYSIS TECHNIQUES IN HIGHER EDUCATION RESEARCH

Dr. Puneet Bhushan
Assistant Professor, Himachal Pradesh University Business School, HPU, Summer Hill, Shimla

ABSTRACT

Multivariate data analysis plays an important role in higher education research. Multivariate data analysis allows researchers to create effective knowledge from load of information which has been collected by the researcher. Multivariate analysis techniques are popular because they enable researchers to create knowledge and thereby improve their analysis as well as interpretation of the data. Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. Multivariate data analys**is** refers to any statistical technique used to analyze data that arises from more than one variable. This essentially models reality where each situation, product, or decision involves more than a single variable. The objective of this paper is to discuss the role of various multivariate data analysis techniques which are used by researchers in the field of higher education. The purpose of this paper is to provide an understanding of various multivariate data analysis techniques, resulting in an understanding of the appropriate uses for each of the techniques. This paper will not provide a discussion of the underlying statistics of each technique whereas it will serve as a guide to understand the types of research questions that can be formulated and the capabilities and limitations of each technique in answering those questions.

INTRODUCTION

Multivariate analysis refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation. Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. This essentially models reality where each situation, product, or decision involves more than a single variable. The information age has resulted in masses of data in every field. Despite the quantum of data available, the ability to obtain a clear picture of what is going on and make intelligent decisions is a challenge. When available information is stored in database tables containing rows and columns, Multivariate Analysis can be used to process the information in a meaningful fashion. Multivariate analysis plays an important role in the understanding of complex data sets requiring simultaneous examination of all variables. Breaking through the apparent disorder of the information, it provides the means for both describing and exploring data, aiming to extract the underlying patterns and structure.

In order to understand multivariate analysis, it is important to understand some of the terminology. A variate is a weighted combination of variables. The purpose of the analysis is to find the best combination of weights. Nonmetric data refers to data that are either qualitative or categorical in nature. Metric data refers to data that are quantitative, and interval or ratio in nature.

Before understanding various multivariate data analysis techniques, it is important to have a clear understanding of the form and quality of the data. The form of the data refers to whether the data are nonmetric or metric. The quality of the data refers to how normally distributed the data are. Another data quality measure is outliers, and it is important to determine whether the outliers should be removed. If they are kept, they may cause a distortion to the data; if they are eliminated, they may help with the assumptions of normality. The key is to attempt to understand what the outliers

represent. The following is the description of various multivariate data analysis techniques which can be used by the researchers in the field of higher education.

MULTIPLE REGRESSION ANALYSIS

Multiple regression is the most commonly utilized multivariate technique. It is an extension of bivariate correlation. It examines the relationship between a single metric dependent variable and two or more metric independent variables with an objective of forecasting the value of dependent variable given the values of independent variables. The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity, and equal variance are carefully observed. The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly. Multiple regression is often used as a forecasting tool.

LOGISTIC REGRESSION ANALYSIS

This technique is a variation of multiple regression where dependent variable is categorical and not metric. If dependent variable has two categories, then binomial logistic regression is used and if dependent variable has more than two categories then multinomial logistic regression is used. The independent variables can be either discrete or continuous. A contingency or classification table is produced, which shows the classification of observations as to whether the observed and predicted events match. The sum of events that were predicted to occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events, is a measure of the effectiveness of the model. This tool helps predict the choices consumers might make when presented with alternatives.

DISCRIMINANT ANALYSIS

Discriminant analysis is used to predict group membership. This technique is used to classify individuals/objects into one of the alternative groups on the basis of a set of predictor variables. The purpose of discriminant analysis is to correctly classify observations or people into homogeneous groups. When there are two groups (categories) of dependent variable, we have two-group discriminant analysis and when there are more than two groups, it is a case of multiple discriminant analysis. The dependent variable in discriminant analysis is categorical and on a nominal scale, whereas the independent or predictor variables are either interval or ratio scale in nature. Discriminant analysis builds a linear discriminant function, which can then be used to classify the observations. The overall fit is assessed by looking at the degree to which the group means differ (Wilkes Lambda or D2) and how well the model classifies. To determine which variables have the most impact on the discriminant function, it is possible to look at partial F values. The higher the partial F, the more impact that variable has on the discriminant function. Multiple discriminant analysis has widespread application in situations in which the primary objective is to identify the group to which an object (e.g. person, firm or product) belongs. Potential applications include predicting the success or failure of a new product, determining the category of credit risk for a person, or predicting whether a firm will be successful. In each instance, the objects fall into groups, and the objective is to predict and explain the bases for each object's group membership through a set of independent variables selected by the researcher.

MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

The extension of univariate analysis of variance (ANOVA) to the involvement of multiple dependent variables is termed as multivariate analysis of variance (MANOVA). This technique examines the relationship between several categorical independent variables and two or more metric dependent variables. Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means and F tests between three or more means), MANOVA examines the dependence relationship between a set of dependent measures across a set of groups. The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be

determined. MANOVA is useful in experimental situations where at least some of the independent variables are manipulated. It has several advantages over ANOVA. First, by measuring several dependent variables in a single experiment, there is a better chance of discovering which factor is truly important. Second, it can protect against Type I errors that might occur if multiple ANOVA's were conducted independently. Additionally, it can reveal differences not discovered by ANOVA tests.

FACTOR ANALYSIS

Factor analysis is a data reduction technique which is used to reduce a large number of variables to a smaller set of underlying factors that summarise the essential information contained in the variables. When there are many variables in a research design, it is often helpful to reduce the variables to a smaller set of factors. Factor analysis is a multivariate statistical technique in which there is no distinction between dependent and independent variables. In factor analysis all variables under investigation are analysed together to extract the underlined factors. Factor analysis is an interdependence technique whose primary purpose is to define the underlying structure among the variables in the analysis. The basic principle behind the application of factor analysis is that the initial set of variables should be highly correlated. If the correlation coefficients between all the variables are small, factor analysis may not be an appropriate technique. Appropriateness of applying factor analysis can be judged by using Kaiser's Measure of Sampling Adequacy (MSA) and Barlettt's test of sphericity.

There are two main factor analysis methods: common factor analysis, which extracts factors based on the variance shared by the factors, and principal component analysis, which extracts factors based on the total variance of the factors. Common factor analysis is used to look for the latent (underlying) factors, whereas principal component analysis is used to find the fewest number of variables that explain the most variance. The first factor extracted explains the most variance.

CLUSTER ANALYSIS

The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects. The division is accomplished on the basis of similarity of the objects across a set of specified characteristics. The resulting clusters should exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. Thus, if the classification is successful, the objects within clusters will be close together when plotted geometrically, and different clusters will be far apart. There are four main rules for developing clusters: the clusters should be different, they should be reachable, they should be measurable, and the clusters should be profitable (big enough to matter). This is a great tool for market segmentation, segmenting industries/sectors, segmenting financial sectors/instruments etc.

MULTIDIMENSIONAL SCALING (MDS)

Multidimensional scaling refers to a series of techniques that help the researcher identify key dimensions underlying respondents' evaluation of objects and then position these objects in this dimensional space. The purpose of MDS is to transform consumer judgments of similarity into distances represented in multidimensional space. This is a decompositional approach that uses perceptual mapping to present the dimensions. As an exploratory technique, it is useful in examining unrecognized dimensions about products and in uncovering comparative evaluations of products when the basis for comparison is unknown. Multidimensional scaling is often used in marketing to identify key dimensions underlying customer evaluations of products, services, or companies. Other common applications include the comparison of physical qualities, perceptions of political candidates or issues, and even the assessment of cultural differences between distinct groups. MDS can infer the underlying dimensions using only a series of similarity or preference judgements about the objects provided by respondents.

CORRESPONDENCE ANALYSIS

This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings. However, unlike MDS, both independent variables and dependent variables are examined at the same time. This technique is more similar in nature to factor analysis. It is a compositional technique, and is useful when there are many attributes and many companies. It is most often used in assessing the effectiveness of advertising campaigns. It is also used when the attributes are too similar for factor analysis to be meaningful. The main structural approach is the development of a contingency (crosstab) table. This means that the form of the variables should be nonmetric. The model can be assessed by examining the Chi-square value for the model. Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.

CONJOINT ANALYSIS

Conjoint analysis is an emerging dependence technique that brings new sophistication to the evaluation of objects, such as new products, services or ideas. Conjoint analysis is often referred to as "trade-off analysis," since it allows for the evaluation of objects and the various levels of the attributes to be examined. It is both a compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed. A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level. Models can be built that identify the ideal levels and combinations of attributes for products and services. The most direct application is in new product or service development, allowing for the evaluation of complex products while maintaining a realistic decision context for the respondent. The market researcher is able to assess the importance of attributes as well as the levels of each attribute while consumers evaluate only a few product profiles, which are combinations of product levels.

CANONICAL CORRELATION

Canonical correlation analysis can be viewed as a logical extension of multiple regression analysis. The most flexible of the multivariate techniques, canonical correlation simultaneously correlates several independent variables and several dependent variables. The underlying principle in canonical correlation is to develop a linear combination of each set of variables (both independent and dependent) in a manner that maximises the correlation between the two sets. This powerful technique utilizes metric independent variables, unlike MANOVA, such as sales, satisfaction levels, and usage levels. It can also utilize nonmetric categorical variables. This technique has the fewest restrictions of any of the multivariate techniques, so the results should be interpreted with caution due to the relaxed assumptions.

STRUCTURAL EQUATION MODELING

Unlike the other multivariate techniques discussed, structural equation Modeling (SEM) examines multiple relationships between sets of variables simultaneously. SEM can examine a series of dependence relationships simultaneously. It is particularly useful in testing theories that contain multiple equations involving dependence relationships. SEM is a family of statistical models that seek to explain the relationships among multiple variables. SEM examines the structure of interrelationships expressed in a series of equations, similar to a series of multiple regression equations. These equations depict all of the relationships among constructs (the dependent and independent variables) involved in the analysis. Constructs are unobservable or latent factors represented by multiple variables. Thus far each multivariate technique has been classified either as an interdependence or dependence technique. SEM can be thought of as a unique combination of both types of techniques. SEM represents a family of techniques, including LISREL, latent variable analysis, and confirmatory factor analysis. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis.

CONCLUSIONS

Each of the multivariate techniques described above has a specific type of research question for which it is best suited. Each technique also has certain strengths and weaknesses that should be clearly understood by the researcher before attempting to interpret the results of the technique. Usually what is happening is higher education research is that researchers are not using the appropriate techniques of data analysis as per the research problem, but they use that technique with which they are familiar or feel comfortable with. The repercussion of doing this is that they do not get the desired results and are also not able to interpret the results in clear and lucid manner. Thus an understanding of various multivariate data analysis techniques is a must for a researcher in the field of higher education.

REFERENCES

Applied Multivariate Data Analysis, Second Edition. (n.d.). Retrieved September 25, 2017, from http://onlinelibrary.wiley.com/book/10.1002/9781118887486.

Chawla, D., &Sondhi, N. (2011). Research methodology: concepts and cases. New Delhi: Vikas Publishing House.

Coakes, S. J., & Ong, C. (2013). SPSS analysis without anguish version 20.0 for windows. Milton: John Wiley.

Decision analyst. (2016, November 14). Eleven Multivariate Analysis Techniques: Key Tools In Your Marketing Research Survival Kit. Retrieved October 5, 2017, from https://www.decisionanalyst.com/whitepapers/multivariate/.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2015). Multivariate data analysis (7th ed.). Pearson Education Limited.

Malhotra, N. K. (2014). Basic marketing research. Pearson Education Limited.

Multivariate analysis. (2017, October 05). Retrieved October 1, 2017, from https://en.wikipedia.org/wiki/Multivariate_analysis.